Homework 3

The purpose of this homework is to practice using the bootstrap to construct confidence intervals, and to learn how to use randomization methods to run hypothesis tests. Please fill in the appropriate code and write answers to all questions in the answer sections, then submit a compiled pdf with your answers through gradescope by 11:59pm on Sunday September 22nd.

As always, if you need help with any of the homework assignments, please attend the TA office hours which are listed on Canvas and/or ask questions on Piazza. Also, if you have completed the homework, please help others out by answering questions on Piazza.

If you want to learn more about the bootstrap (and why it is useful for teaching the concept of confidence intervals), please read this paper by Tim Hesterberg

Problem 1: Calculating confidence intervals using the bootstrap

As discussed in class, we can use the bootstrap to estimate standard errors which can then be used to calculate confidence intervals. Let's use the bootstrap to calculate a confidence interval for the mean height of OkCupid users.

Part 1.1 (10 points)

To explore how confidence intervals work when a sample size is relatively small (and also to make this problem more computationally efficient), use the heights from only the first 20 OkCupid users, and then do the following steps:

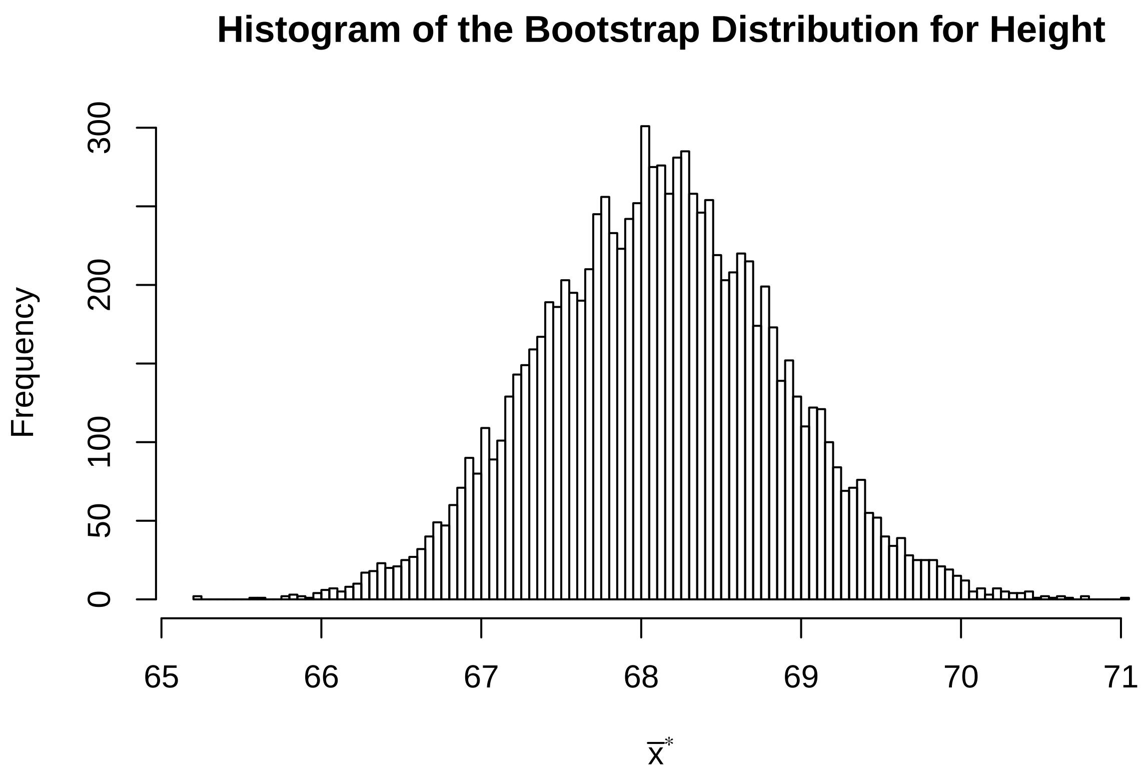
1) Estimate the standard error of the mean height using the bootstrap
2) Plot a histogram of the bootstrap distribution
3) Calculate an approximate 95% confidence interval for the heights of OkCupid users using the formula CI $[\bar{x} - 2 \cdot SE^*, \bar{x} + 2 \cdot SE^*]$

Report your confidence interval below and describe what the confidnece interval tells you.

```r
# load the OkCupid data and extract the heights for the first 20 profiles
library(okcupiddata)
the_heights <- profiles$height[1:20]


# construct the bootstrap distribution
bootstrap_dist <- NULL
for (i in 1:10000){
  boot_sample <- sample(the_heights, replace = TRUE)
  bootstrap_dist[i] <- mean(boot_sample)
}
```

```r
# plot a histogram of the bootstrap distribution
hist(bootstrap_dist,
     nclass = 100,
     xlab = TeX("$\\bar{x}$^*"),
     main = "Histogram of the Bootstrap Distribution for Height")
```



```r
# calculate the standard error
(SE <- sd(bootstrap_dist))
```

```
## [1] 0.7496106
```

```r
# calculate 95% confidence intervals
CI_lower <- mean(the_heights) - 2 * SE
CI_upper <- mean(the_heights) + 2 * SE

c(CI_lower, CI_upper)
```

```
## [1] 66.65078 69.64922
```

Answer:

The 95% confidence interval is [66.64, 69.66]. This means that there is a 95% chance that this interval contains the true population mean. The standard error SE = 0.749.

Part 1.2 (10 points)

Run your code above again but 1) use the first 100 OkCupid users, and 2) then use the first 1000 OkCupid users. Report what the confidence interval are when using these different number of users (i.e., when using different sample sizes n). Do they seem much smaller? Note: you do not need to show code here, just modify the code above rerun it and then report the results.

Answer:
The 95% confidence interval using the first 100 users is [68.17, 69.63] (Standard error = 0.363). Using the first 1000 users the interval is [68.16, 68.66] with a standard error of 0.125. The interval decreases in size as the sample size increases.

Part 1.3 (10 points)

Now write code to create confidence intervals separately for the heights of male and female OkCupid users by using the subset() function to get separate vectors of heights for males and females and use the first 100 male and 100 female okcupid users. Does it appear plausible that the actual mean height for males μ_{male} is the same as the actual mean height for females μ_{female} ?

```r
# get the hights for the first 100 male and 100 female OkCupid users
the_heights_male <- subset(profiles, sex == "m")

# continue from here...
hundred_male <- the_heights_male$height[1:100]

the_heights_female <- subset(profiles, sex == "f")
hundred_female <- the_heights_female$height[1:100]



# create bootstrap distributions for the male and female heights
bootstrap_dist_m <- NULL
for (i in 1:10000){
  boot_sample_m <- sample(hundred_male, replace = TRUE)
  bootstrap_dist_m[i] <- mean(boot_sample_m)
}

bootstrap_dist_f <- NULL
for (i in 1:10000){
  boot_sample_f <- sample(hundred_female, replace = TRUE)
  bootstrap_dist_f[i] <- mean(boot_sample_f)
}


# calculate standard errors and confidence intervals for the male and female heights
(SE_m <- sd(bootstrap_dist_m))
```

```
## [1] 0.2715646
```

```
(SE_f <- sd(bootstrap_dist_f))
```

```
## [1] 0.2444839
```

```
# calculate 95% confidence intervals
(CI_male <- c((mean(hundred_male) - 2 * SE_m),
              (mean(hundred_male) + 2 * SE_m)))
```

```
## [1] 69.99687 71.08313
```

```
(CI_female <- c((mean(hundred_female) - 2 * SE_f),
              (mean(hundred_female) + 2 * SE_f)))
```

```
## [1] 64.56103 65.53897
```

Answers: The 95% Confidence Interval for male heights is [69.99, 71.09] with a standard error of 0.276. The 95% Confidence Interval for female heights is [64.56, 65.54] with a standard error of 0.276. Since these intervals do not overlap, it does not seem plausible that $\mu_{male} = \mu_{female}$.

Problem 2: Comparing bootstrap CIs to formula based CIs

As you (almost certainly) learned in Introduction to Statsistics, there is a mathematical forumla that can give you the standard error for the mean statistic $\bar{x}$ (note: standard error for the mean statistic is called "the standard error of the mean" and is often abbreviated SEM or denoted as $\sigma_{\bar{x}}$ or $s_{\bar{x}}$).

The formula for the SEM is:

(1) $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

and an estimate for this is given by:

(2) $s_{\bar{x}} = \frac{s}{\sqrt{n}}$

Where:

- σ is the population standard deviation,
- s is the standard deviation statistic computed from a sample of size n
- n is the sample size.

Note, equation 1 above is a theoretical construct since we will never know σ (only Plato knows this) while equation 2 above is possible to calculate from a sample of data.

Part 2.1 (10 points)

Using the formula in equation 2 above, repeat the analyses in problem 1.1 by calculating the standard error of the mean, and a 95% confidence interval for the mean height of OkCupid users, but use formula 2 to caluclate the standard error rather than the bootstrap. Again, use only the first 20 OkCupid users in the data set. Is the confidence interval you created using formula 2 close to the confidence interval you created in problem 1.1?

```r
# calculate the SEM and CI using the formula in equation 2
the_heights <- profiles$height[1:20]
(SE_eq <- (sd(the_heights))/sqrt(20))
```

```
## [1] 0.7687139
```

```r
(CI_eq <- c((mean(the_heights) - 2*SE_eq),
             (mean(the_heights) + 2*SE_eq)))
```

```
## [1] 66.61257 69.68743
```

Answers: The 95% confidence interval is [66.61, 69.69] with a standard error of 0.769. The mean remains the same, but the standard error has increased by 0.02, resulting in a wider interval.

Part 2.2 (15 points)

The line of code below extracts the incomes from the first 10 OkCupid users. Compare the confidence intervals for the mean income using:

1) the formula for the standard error
2) the bootstrap estimate of the standard error
3) the bootstrap percentile method

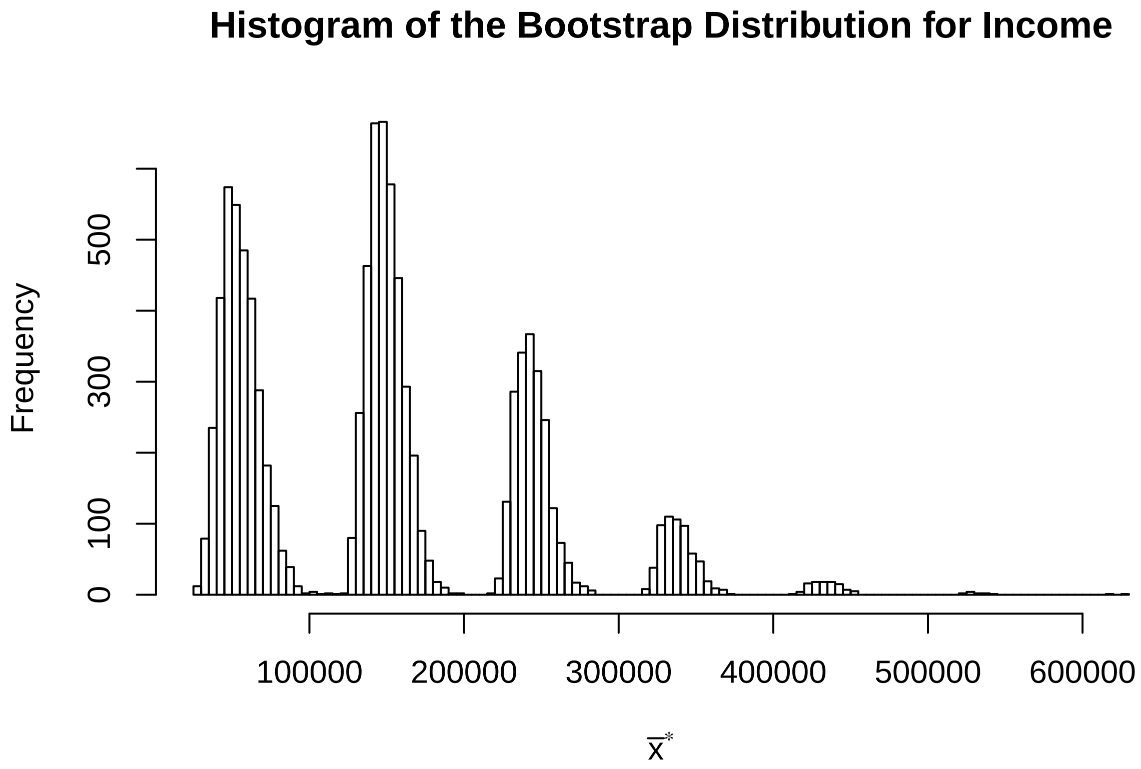
Report which confidence interval(s) seems to give the most reasonable results, and give an explanation for why they differ. Also, assuming that the whole population is just the OkCupid users in the profiles data frame, calculate the value of the parameter that these confidence intervals are trying to capture (using the appropriate symbol) and report whether all these methods capture this parameter value.

```r
# extract OkCupid user's incomes from the users that listed their income
the_income <- na.omit(profiles$income)[1:10]


# create a bootstrap distribution
bootstrap_dist_income <- NULL
for (i in 1:10000){
  boot_sample_income <- sample(the_income, replace = TRUE)
  bootstrap_dist_income[i] <- mean(boot_sample_income)
}

# plot the boostrap distribution

hist(bootstrap_dist_income,
     nclass = 100,
     xlab = TeX("$\\bar{x}$^*"),
     main = "Histogram of the Bootstrap Distribution for Income")
```



```
# calculate the bootstrap estimate of the standard error SE*
(SE_inc <- sd(bootstrap_dist_income))
```

```
## [1] 90315.51
```

```
# calculate a CI using the bootstrap perentiles
(CI_boot_percentile <- quantile(bootstrap_dist_income, c(.025, .975)))
```

```
##   2.5%  97.5%
##  39000 346000
```

```
# calculate the CI based on using equation 2 to estimate the SE
(CI_boot_SE <- c(mean(the_income) - 2 * SE_inc, mean(the_income) + 2 * SE_inc))
```

```
## [1] -30631.02 330631.02
```

```
# calculate the parameter value
mean(the_income)
```

```
## [1] 150000
```

```r
#calculate using the fomula
(SE_form <- sd(the_income)/sqrt(10))
```

```
## [1] 95207.38
```

```r
(CI_boot_formula <- c(mean(the_income) - 2 * SE_form, mean(the_income) + 2 * SE_form))
```

```
## [1] -40414.75 340414.75
```

Answers

As we can see from the histogram of the bootstrap distribution, there are multiple peaks at which a particular sample mean value will occur at most often. This means the original sample contains a wide distribution of values, and since the sample size is small, the value of the mean can vary drastically depending on which elements of the sample are chosen. This also means the sample error of the distribution will be very large, which causes the confidence interval to be too large, containing negative values. The problem with the bootstrap distribution and formula methods is that they assume the sample is normally distributed, so the 95% confidence interval will hold. However, our sample is not normally distributed and thus it is more logical to use the bootstrap percentile, which makes calculations based off the values which are actually in the set.

Reflection (5 points)

Please reflect on how the homework went by going to Canvas, going to the Quizzes link, and clicking on Reflection on homework 3